

Grammar rules for the isiZulu complex verb

C. Maria Keet¹ and Langa Khumalo²

¹ Department of Computer Science, University of Cape Town, South Africa,
mkeet@cs.uct.ac.za

² Linguistics Program, School of Arts, University of KwaZulu-Natal, South Africa,
khumalol@ukzn.ac.za

Abstract. The isiZulu verb is known for its morphological complexity, which is a subject for on-going linguistics research, as well as for prospects of computational use, such as controlled natural language interfaces, machine translation, and spellcheckers. To this end, we seek to answer the question as to what the precise grammar rules for the isiZulu complex verb are (and, by extension, the Bantu verb morphology). To this end, we iteratively specify the grammar as a Context Free Grammar, and evaluate it computationally. The grammar presented in this paper covers the subject and object concords, negation, present tense, aspect, mood, and the causative, applicative, stative, and the reciprocal verbal extensions, politeness, the wh-question modifiers, and aspect doubling, ensuring their correct order as they appear in verbs. The grammar conforms to specification.

Keywords: Bantu languages, isiZulu, Spell-checking, Verb

1 Introduction

While South Africa recognises eleven official languages, only English and Afrikaans have significantly invested in computational resources. IsiZulu, which is the most widely spoken language in South Africa, still remains under-resourced. In this article we focus on the development of perfect grammar rules for the isiZulu verb and by extension Bantu verb morphology. It is notable that a small Definite Clause Grammar and POS tagger for isiZulu has been proposed and is available online [27]. It covers only a fraction of the complexities of the isiZulu verb; for instance, it addresses only one extension to the exclusion of the causative, applicative, and the reciprocal extensions. Other attempts to formal approaches to isiZulu morphology focus predominately on nouns rather than verbs [23,24], or describe only a few sample regular expressions that cover a very small fraction of the verb [3]. The morphology of the verb is widely regarded as the most interesting theoretically. Sections 2 and 3 provides a brief discussion on this interesting grammatical category whose complexity presents challenges to the computation and generation of grammar rules. Traditional accounts on isiZulu grammar are based on dated sources [5,6] and limited accounts on Wikipedia. There is no comprehensive synchronic grammar of isiZulu yet.

We present a morphological analysis of the isiZulu verbal extension and rules for that. This is done in order to create a spell checking and part-of-speech tagging of the verb in isiZulu. We explore the means to automate the checking of the complex verb morphology. We ultimately address the following question: *What are the precise grammar rules for the isiZulu verb (and, by extension, the Bantu verb morphology)?* We thus formalise the grammar for the isiZulu verb as a Context-Free Grammar. This grammar is subsequently represented computationally so as to test its correctness with respect to specification, using a set of words and generating their derivations in the JFlap tool. The grammar covers not only the usual subject and object concords, but also negation, present tense, aspect, mood, and the verbal extensions such as the causative, applicative, stative and the reciprocal, politeness, the wh-questions modifiers, and aspect doubling.

The paper is structured as follows. Section 2 gives a synchronic outline of the isiZulu verb morphology and also highlights comparative salient features that are characteristic of Bantu languages and Section 3 discusses related works. The main contribution, the formalised account of the isiZulu verb for present tense, is presented in Section 4 and evaluated in Section 5. We conclude in Section 6.

2 Basics of the isiZulu verb

IsiZulu is a Bantu language that belongs to the Nguni³ group of languages. It has close affinity to other Nguni language varieties. Bantu languages have a characteristically agglutinating morphology, which makes their structure rich and complex. The agglutinating typology is not unique to Bantu languages as other agglutinating languages with extremely complex morphology include Turkish, Hungarian, and Finnish [7]. In characterising the complexity of the verbal constructions in Bantu languages, [28] (p291) states that the morphology of the verb shows “[...] the fullest extent of the agglutinative nature of the Bantu language family”. Such complex morphology presents a lot of challenges in attempts to develop computational technologies in isiZulu.

The isiZulu verbal morphology typically comprise of a verb root (VR) to which extensions such as the causative, applicative, reciprocal, passive etc. are suffixed and to which morphemes that encode negation (NEG), subject marker (SM) and object marker (OM) that cross-reference noun phrases (NPs), tense/aspect, modality, etc. are prefixed.

At the core of the verbal structure is a root morpheme, which is called the verb root (VR). The VR forms the nucleus of the verbal morphology. This core element supports a number of affixes, both prefixes and suffixes. Each affix type occupies a specific position in the verbal morphology. The affixes include the SM, the OM, Tense Aspect and Mood (TAM), and various derivational extensions. The verb is characteristically terminated with a final vowel (FV) and this final

³ A term used by Guthrie [10] to classify isiZulu, isiXhosa, isiNdebele and siSwati in Group S, Zone 40.

vowel of the verb may encode mood, tense, polarity and potential modality. Fig. 1 illustrates the complex verb in Bantu.

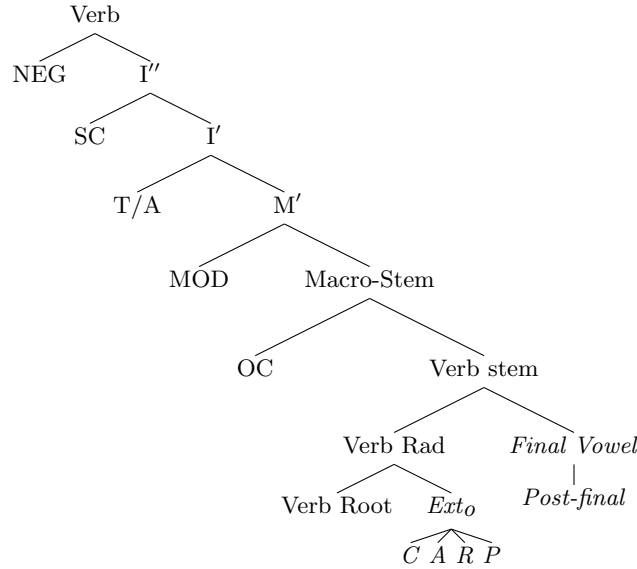


Fig. 1. The structure of a complex verb in Bantu, where the elements not in italics font are considered to be the canonical verb structure. NEG: negative; SC: subject concord; T/A: tense/aspect; MOD: mood; OC: object concord; Verb Rad: verb radical; C: causative; A: applicative; R: reciprocal; P: passive.

Khumalo [15] (p79) proposes a verb slot system for the complex verbal form⁴ in Ndebele, which is applicable to isiZulu, as included in Table 1. The prefixed morphemes differ from suffixed extensions in both form and function. Formally the suffixes have a -VC- structure, as opposed to the regular CV syllable structure. Functionally the verbal extensions affect the argument structure [19] (p203). Example (1) shows the morphological organisation of the verb in isiZulu.

- (1) *Aba-shana* *ba-ya-zi-theng-is-el-an-a* *izimpahla*
 2.Children 2SM-Pres-8OM-buy_{VR}-CAUS-APPL-REC-FV 8.clothes
 ‘The children are selling the clothes to each other’

The VR *-theng-* ‘buy’ supports the extensions *-is-* for the causative, *-el-* for the applicative, *-an-* for the reciprocal, and the prefix clitics *ba-* for the ‘subject marker’, *-ya-* for the ‘tense’, and *-zi-* for the ‘object marker’.

⁴ The following abbreviations are used: A=aspect; ADV=adverb; APPL=applicative; CONT=continuous tense; EXCL=exclusive aspect; Ext=extension; FV=final vowel; M=mood; NEG=negative tense; OC=object concord; PROG=progressive tense; Rad=radical; SG=singular; SC=subject concord; T=tense; VR=verb root; VS=verb stem

Table 1. Bantu verb slot template, adapted from [20]; ~: also realised as

Slot	Pre-initial	Initial	Post-initial	Pre-radical	Radical	Pref-final	Final	Post-final
<i>Function</i>	TAM, NEG, clause type	SM	TAM, NEG, SM	OM	VR	TAM, valence change (CARP)	FV	Participant, NEG, clause type
<i>Example</i>	a	ngi	za, nga	ba	khal	is (el; an; w)	a	(ni ~ nini) ¹

¹ The plural suffixes denote both general plurality and honorific plurality.

The verb extensions interact in complex ways with the valency of the base verb. The extensions for several languages are listed in Table 2. Semantically (with the exception of the passive extension) they alter the number of participants expressed by the verb. Grammatically they alter the number of arguments present expressed by an NP or a pronominal element.

Table 2. Verbal extensions in Proto Bantu, Swahili, isiZulu and isiNdebele. *: morphemes belong to an ancestor language or proto form, v: precise phonetic form of the vowel. (Adapted from [26], p72.)

Derivational Extension	Proto Bantu	Swahili	Zulu	Ndebele
causative	*-i-/ -ici-	-ish-, -esh-	-is-	-is-
applicative/dative	*-il-	-i-, -e-	-el-	-el-
reciprocal/associative	*-an-	-an-	-an-	-an-
passive	*-v-/ -ibu	-w-	-iw-, -w-	-iw-, -w-
stative/neutro-passive/positional	*-am-	-ik-, -ek-	-ek-	-ek-, -akal-
reversive/separative	*-ul-; uk-	-u-, -o-	-ul-	-ul-, -ulul-, -uk-
neuter	*-ik-		-akal-	
extensive	*-al-			-isis-
repetitive	*-ag- ~ -ang-			
impositive	*-ik-			
tentive/contative	*-at-			

2.1 Concordial agreement

The term agreement in Bantu is often used alongside the term concord. These two terms are sometimes used interchangeably. Agreement occurs when grammatical information appears on a verb which typically is not the source of that information. This is done through a series of agreement markers called concords that are affixed to the verb. The noun or pronoun is said to govern the agreement of all words associated with it in a syntactical relationship [29] (p8). Agreement

Table 3. Basic verb conjugation. NC: noun class; SC: subject concord; OC: object concord; NEG SC: negative subject concord.

Conjugation for noun classes				Conjugation for persons			
NC	SC	NEG SC	OC	Pers. Pron.	SC	NEG SC	OC
1	u-	aka-	-m-	I	ngi-	angi-	-ngi-
2	ba-	aba-	-ba-	you (sing.)	u-	awu-	-ku-
1a	u-	aka-	-m-	he/she	u-	awu-	-m-
2a	ba-	aba-	-ba-	we	si-	asi-	-si-
3a	u	aka-	wu	you (pl.)	ni-	ani-	-ni-
(2a)	ba-	aba-	-ba-	they	ba-	aba-	-ba-
3	u-	awu-	-wu-				
4	i-	ayi-	-yi-				
5	li-	ali-	-li-				
6	a-	awa-	-wa-				
7	si-	asi-	-si-				
8	zi-	azi-	-zi-				
9a	i	ayi-	yi				
(6)	a-	awa-	-wa-				
9	i-	ayi-	-yi-				
10	zi-	azi-	-zi-				
11	lu-	alu-	-lu-				
(10)	zi-	azi-	-zi-				
14	bu-	abu-	-bu-				
15	ku-	aku-	-ku-				

is thus a cross-referencing device for subjects and objects. Table 3 shows the conjugation of the verb in isiZulu for all the noun classes and persons. As shown in the table the verb not only takes the subject and object concords, but also the negative subject concord.

As stated above, the verbal structure consists entirely of bound morphemes. These are the VR and a number of affixes such as the subject concord (SC), the object concord (OC), Tense Aspect Mood (TAM), and various other derivations (CARP), typically terminated with a final vowel (FV). The example below is a Chishona verb *ndichaenda* ‘I will go’

(V2) *ndi - cha - end - a* Chishona: *ndichaenda*
1.SC - 1.TM - Root - FV
‘I’ ‘will’ go ‘I will go’

2.2 More on clitics of the verb

The elements prefixed to the verb stem in isiZulu are usually referred to as clitics. Clitics are independent syntactic elements which appear as part of the host word. This independent element is involved in a morphological merger

to appear phonologically as part of a derived word. Example 3 is illustrative:

(V3) *ngi - m - bon - a kusasa Ngimbona kusasa.*

1.SC - 1.OC - Root - FV

‘I’ ‘him’ see tomorrow ‘I (will) see him tomorrow.’

The independent elements *ngi-* and *m-* merge to form a derived word *ngimbona*. The clitics are thus syntactic elements, which lack phonological independence. They cannot stand or appear on their own. It is clear that syntactically they are words but phonologically they are not. They are not viewed as phonological words because they fail to satisfy the minimality condition for being a word in Bantu. The Bantu condition is that a word has to minimally consist of two syllables. In the example above, *ngi-* is a single syllable and *m-* is also a single syllable known as “syllabic *m*”. The notion of clitics and their grammatical status in Bantu is still a very interesting one.

The isiZulu verbal suffixes are also bound morphemes without any independent status, hence they are also clitics. They are involved in the determination of expressible NP arguments within the sentence. As stated earlier, these include the morphology for encoding the causative, applicative, reciprocal, passive, stative, etc. These suffixes, together with the VR, are terminated by the FV *-a* and together make up the verb stem (VS) as shown in Figure 1. The following example shows the VR plus the verbal extensions.

(V4) *bon - a bona* ‘see’ un-extended verb

VR - FV

bon - is - a bonisa ‘make see’ extended verb

bon - el - a bonela ‘see for’ extended verb

bon - an - a bonana ‘see each other’ extended verb

While the suffixes (or verb extensions) clearly introduce a new syntactic element, they however are themselves not independent. They cannot stand as phonological words on their own, hence, they are clitics. The clitics in Bantu can co-occur with the verbal extensions. However, when this happens, they are attached outside the final vowel. The extensions appear to be more intimately connected to the host VR. Crucially, while the VS is the domain of a number of linguistic processes, its influence is not extended to the suffixed clitics. It is thus assumed that the VS has lexical integrity. This makes the VS an important subdomain in the morphological structure of the verb. The VS is thus the domain of lexical processes in Bantu [18].

2.3 Aspects of isiZulu Tense

Bantu languages typically consist of rich tense and aspect systems, characterised by various temporal distinctions [16]. The complexity of grammaticalised tense and aspect in isiZulu is exemplified by its five tenses. The tenses include the remote past, recent past, present, immediate future and remote future tense. The three aspectual forms are the simple, progressive and exclusive aspect.

IsiZulu makes productive use of its grammatical aspect system. The Progressive aspect in isiZulu is denoted by the affix *-sa-*. Whilst conveying an ongoing

action/state/event, the morpheme also carries an inherent adverbial meaning of ‘still’ as shown in the example below.

Ngi-**sa**-fund-a isiZulu
 1SG-PROG-VR-FV isiZulu
 ‘Even now I am still studying Zulu’

There is no direct adverb (lexical item) for the English word ‘still’ in isiZulu. Instead it is expressed using the adverb *namanje*, which directly translated means ‘even now’. The rich expression of temporal events and situations in isiZulu, is further highlighted in the following example.

Na-manje ngi-**sa**-fund-a isiZulu
 Cl-ADV 1SG-PROG-VR-FV isiZulu
 ‘I am still studying isiZulu’

Similarly, the Exclusive morpheme *se-* expresses an inherent adverbial aspect, meaning ‘now’. This morpheme may be used with the adverb *manje* ‘now’, thereby expressing double aspect comprising of the grammatical aspect (*se-*, now) + grammatical aspect (*manje*, ‘now’). This phenomenon has been referred to as aspect doubling, and is illustrated below.

Se-ngi-ya-fund-a manje
 EXCL-1SG-CONT-VR-FV ADV
 ‘Now, I am now studying’

The productive nature of Exclusive and Progressive aspect morphemes in isiZulu has not received considerable attention. The Exclusive morpheme *se-* may be used with adverbial structures conveying similar meanings in isiZulu, while this is proscribed in the English language.

Section 2 has thus shown the complexity of the morphology of the verb. It has shown that not only does isiZulu verb get inflected before the VR but also after the VR through a whole gamut of clitics that have an effect on the construction of a whole sentence. This is not unique to isiZulu but is characteristic of other Bantu languages like Chishona. It is thus this complexity of verbal morphology which presents challenges in the development of computational technologies in isiZulu.

3 Related work on isiZulu verbs

The verb in Bantu has received considerable attention (cf. [11,17,18] etc.). This is because it is arguably the most interesting grammatical category in linguistic theory. Many accounts in Bantu have sought to explicate the many salient morphosyntactic properties of the verb using different generative theoretical approaches. Buell [4] is the most recent comprehensive study of isiZulu verb. Buell discusses the isiZulu verb using a restrictive theory of syntax, which is premised on the assumption that there is a close relation between the morphology and

the syntax. His account covers an array of inflectional elements such as mood, sub-mood, and polarity, subject and object agreement. Buell also makes reference to, albeit briefly, the verbal suffixes such as the applicative [4]. In a study such as his, it is impossible to be exhaustive. In this study, however, we cover the causative, applicative, reciprocal and the passive. Earlier studies on the Zulu verb are [1], whose study focuses on the verb and its conjugation of various subject concords and their allomorphs, tense and mood conjugational morphemes. Beuchat [1] does not make reference to derivational extensions.

As we seek to have a precise, formal, representation of the isiZulu verb, we also consider computational processing of the verbs, for they require a structured representation to work computationally. Regarding controlled natural languages and natural language generation, there are only two recent papers [13,14], which cover verbs only to the extent of noun class-appropriate singular present tense when verbalising simple existential quantification involving object properties. Some literature on computational linguistics for isiZulu exists that is relevant to some extent, being morphological analysers. Among these works, the Ukwabelana corpus and related materials [27] is most comprehensive and is the only one with online source material. Besides the corpus and limited semi-automated POSTagging, Spiegler et al. developed a basic Definite Clause Grammar (DCG)⁵, of which a relevant section is shown in Fig. 2. The first to note is that while it

```
v --> neg, spfn, asp, opf, vr1, vsf_neg.
v --> neg, spfn, asp, vr1, vsf_neg.
...
v --> spfi, asp, opf, vr1, vsf.
v --> spfi, asp, vr1, vsf.
...
v --> spfp, vr1, vsf.
...
v --> spfs, opf, vr1, vs.
...
vr1 --> vr, xa.
vr1 --> vr, xc.
vr1 --> vr, xn.
vr1 --> vr, xp.
vr1 --> vr, xr.
vr1 --> vr.
```

Fig. 2. Selection of DCG statements from the online supplementary material to [27] (“...” means line(s) omitted here).

has each of the “CARP” (xc etc.; bottom part), it has only ever one of them. This constitutes a subset of the possibilities, as multiple ones can be appended

⁵ Available from: <http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/resources.jsp>; last accessed on August 19, 2015.

and as they appear in a certain order. Also, the passive (**xp** in the CFG above), which causes changes in the concords in the verb, is not catered for, nor are the politeness prefixes (*aw-*, a.o.) and tenses other than present tense, nor imperative. That is, it covers a subset. That said, it is already useful and at least it can be extended, unlike related works such as [3,23,24]. Bosch and Eiselen [3] report on a basic spelling checker that is based on a set of regular expressions. They illustrate 4 examples that show a few permutations for a verb, e.g.,

`/^(ba)(ya)?(ngi)?(.+)(el)?(a|c)(ni|phi)?$/`

which is a subset of the conjugation (*ba-* for 3rd person plural) and CARP (*-el-*) and no details of its implementation is provided [3]. A related work on morphological rules focuses on nouns [23]. The bootstrapping approach presented in [24] considers the copulative (and a few other word categories) but not verbs in general. Assuming that the *lexc* and *xfst* rules as described in [22] do exist, then its coverage of verb features is incomplete, notably missing mood and aspect, applicative, reciprocal, stative, politeness, and *wh*-ending. While their approach of figuring out which CARP extensions are permitted with a verb root is interesting (relying on the noun forms), it results in rules that are too restrictive: “by explicitly listing the noun stems of the verb root *-fund-* no suffixes other than *-a*, *-el-o*, *-i*, *-is-an-o*, *-is-i*, *-is-o*, *-is-wa*, and *-o* will occur with *-fund-*.” (emphases omitted) [22], but words such as *awufunde* ‘[could we/you] please study’ and *usafundaphi* ‘where are you [still] studying?’ are valid verb forms.

Concerning verbs in other Bantu languages, several rules for Setswana (also an official language in South Africa) verbs have been implemented in *xfst* [21], but it is not clear how much of the grammar of the verb was covered. Further afield from the languages in South Africa, there are exploratory results for Ekegusii (a Bantu language spoken in Western Kenya) with several regular expressions in *xfst* zooming in on the difficulties of tone in relation to verbs [8], and there is a systematic account of the Runyakitara (a Bantu language spoken in Uganda) verb implemented in *fsm2*, including both the grammar and context-dependent rewriting rules that handle morpho-phonological and orthographical issues [12].

From a scientific methodological viewpoint, there is no clear ‘winner’ between the data-oriented approach and the knowledge and rules-based approach to obtain the grammar; or the empirical and the rational paradigms. The data-based techniques, notably machine learning [27,9], have the hurdle of finding or creating a representative enough corpus and at least some rules to process them, whereas the rules-based techniques face the issue of a dearth of up-to-date, structured, grammar books, having to start afresh with formalising the grammar as grammar or regular expressions. Our literature survey indicates the latter approach is used considerably more often for Bantu languages [23,8,12,22,21]. However, use/preference does not imply more effective.

4 Structured representation of the isiZulu verb

Methodologically, theoretically, and technically, there are multiple ways of specifying the grammar of a POS category; e.g., using a grammar such as a DCG,

regular expressions, or their more abstract representation with an automaton (PDA for a CFG). While for the small subset of prefixes for noun classes and some simple verb forms it certainly is easier to design an NFA, transform it into a DFA and from there into a RE, there are so many options with the verbs that the automaton would become too large and wieldy. Moreover, the cross-dependencies of elements before and after the verb root indicates that a regular expression is not expressive enough and may need a CFG rather than an RG. To create the structured representation of the isiZulu verb that is computationally useful, we build it up stepwise from a linguistic pattern, to some quasi regular expressions that in turn revealed a pattern, and from there to a basic grammar, which in turn was extended with other verb features. For reason of exposing this incremental methodological approach to the design of the grammar, we report on the component-steps of one cycle, and subsequently only the outcome of the subsequent cycles, which amount to extensions of the grammar obtained in the first round. The additions to the first cycle were—and can be—done in arbitrary order.

4.1 First iteration

From the general linguistic structure of the isiZulu verb as depicted in Fig. 1, we obtain the full set of ‘slots’ of the verb’s basic components as follows:

R0: [NEG] [SC] [T/A] [MOD] [OC] [VR] [C] [A] [R] [P] [FV]

with [VR] being the verb root at the centre. Each NEG, SC etc. has its own set of characters for each noun class; see Table 3. For the CARP, we have, as a general rule, C = *is*, A = *el*, R = *an*, and for P = *w*, though there is some phonological conditioning for A and P.

First part before the VR Lets consider first what comes before the verb root (VR), with the subject present and active, and both in the positive (thus FV=*a*) and in the negative (FV=*i*), and assuming there is an object after the verb, so that OC can be omitted (see below for OC inclusion). Then the following patterns are permissible (italicised):

- *[SC] [VR] [FV=a]*
- *[SC] [MOD] [VR] [FV=a]*
- *[SC] [T/A] [MOD] [VR] [FV=a]*
- *[NEG] [SC] [VR] [FV=i]*
- *[NEG] [SC] [MOD] [VR] [FV=i]*
- *[NEG] [SC] [T/A] [MOD] [VR] [FV=i]*

This can be captured by the following two quasi regular expressions (where the NEG, SC, T/A, MOD, and VR are to be replaced by the actual strings):

R1: [SC] [T/A]^{0..1} [MOD]^{0..1} [VR] a

R2: [NEG] [SC] [T/A]^{0..1} [MOD]^{0..1} [VR] i

Or, if the software to implement it allows for REs+rules, then:

R3: [NEG]^{0..1} [SC] [T/A]^{0..1} [MOD]^{0..1} [VR] [FV]

R4: if NEG then FV=i, else FV=a

The OC is used if there is no explicit object named after the verb. Then we have the following options:

- [SC] [OC] [VR] [FV=a]
- [SC] [MOD] [OC] [VR] [FV=a]
- [SC] [T/A] [MOD] [OC] [VR] [FV=a]
- [NEG] [SC] [OC] [VR] [FV=i]
- [NEG] [SC] [MOD] [OC] [VR] [FV=i]
- [NEG] [SC] [T/A] [MOD] [OC] [VR] [FV=i]

This amounts to the following two rules

R5: [SC] [T/A]^{0..1} [MOD]^{0..1} [OC] [VR] a

R6: [NEG] [SC] [T/A]^{0..1} [MOD]^{0..1} [OC] [VR] i

While we could combine R1, R2, R5 and R6, it then will have to go through a whole set of permutations to either check correct syntax or generate it. We currently expect it to be quicker to look ahead to the tag of the next phrase to determine whether an OC is needed, and then choose either the rules with OC or without; that is:

R7: if next word==∅ or next word != noun then use R5 or R6, else use R1 or R2

where the “next word==∅” essentially means that the verb is the last word in the sentence.

Second part after the VR The extension is added to the verb root (VR), and comes before the FV. We show a section of the rather long list of all options:

- [some prefix] [VR] [C] [FV]
- [some prefix] [VR] [C] [A] [FV]
- [some prefix] [VR] [C] [A] [R] [FV]
- [some prefix] [VR] [C] [A] [P] [FV]
- [some prefix] [VR] [C] [R] [FV]
- [some prefix] [VR] [C] [R] [P] [FV]
- [some prefix] [VR] [C] [P] [FV]
- [some prefix] [VR] [C] [A] [R] [P] [FV]
- [some prefix] [VR] [A] [FV]
- etc.

That is, the CARP stay in that order, but any one or more of them can be used, so the following quasi regular expression can be specified:

R8: [some prefix][VR] [C]^{0..1} [A]^{0..1} [R]^{0..1} [P]^{0..1} [FV]

to be implemented by filling in the actual strings in the places of the VR, C, A, R, and P, and the [some prefix] following the rules as outlined above.

From quasi RE to grammar The quasi REs show some repetition, and especially the “[*some prefix*]” makes it look clumsy. It also can be seen there are four components: what comes before the VR, the VR, what comes after the VR, and the final vowel. This can be addressed more easily and succinctly with a generative grammar. To design that, let us first convert R1, R2, R5 and R6 into grammar notation, using the following abbreviations: v=verb (with its adornments), n=negation, s=subject concord, t=tense, asp=aspect, o=object concord, m=mood, c=causative, a=applicative, r=reciprocatative, p=passive, vr=verb root, text in true type font are terminals, and spaces in the rules are not spaces in the word, but added for readability:

%%R1 in CFG notation

$v \rightarrow s \text{ vr } a \mid s \text{ m vr } a \mid s \text{ t m vr } a \mid s \text{ asp m vr } a$

%%R2 in CFG notation

$v \rightarrow n \text{ s vr } i \mid n \text{ s m vr } i \mid n \text{ s t m vr } i \mid n \text{ s asp m vr } i$

%%R5 in CFG notation

$v \rightarrow s \text{ o vr } a \mid s \text{ m o vr } a \mid s \text{ t m o vr } a \mid s \text{ asp m o vr } a$

%%R6 in CFG notation

$v \rightarrow n \text{ s o vr } i \mid n \text{ s m o vr } i \mid n \text{ s t m o vr } i \mid n \text{ s asp m o vr } i$

This still will result in duplications, for these will have to be reused for CARP. To this end, we create *pre* and its negated variant *npre* and a *post* (that can be empty, ϵ), that will surround the verb root.

$pre \rightarrow s \mid s \text{ m} \mid s \text{ t m} \mid s \text{ asp m} \mid s \text{ o} \mid s \text{ m o} \mid s \text{ t m o} \mid s \text{ asp m o}$

$npre \rightarrow ns \mid ns \text{ m} \mid ns \text{ t m} \mid ns \text{ asp m} \mid ns \text{ o} \mid ns \text{ m o} \mid ns \text{ t m o} \mid ns \text{ asp m o}$

$post \rightarrow c \mid c \text{ a} \mid c \text{ a r} \mid c \text{ a p} \mid c \text{ r} \mid c \text{ r p} \mid c \text{ p} \mid c \text{ a r p} \mid a \mid a \text{ r} \mid a \text{ r p} \mid a \text{ p} \mid r \mid$
 $r \text{ p} \mid p \mid \epsilon$

This is then put together with the verb root and final vowel:

$v \rightarrow pre \text{ vr } post \text{ a} \mid npre \text{ vr } post \text{ i}$

Let us now complete the grammar so far with the terminals.

1. List of subject concords and negative sc:

$s \rightarrow \text{ngi} \mid \text{u} \mid \text{si} \mid \text{ni} \mid \text{ba} \mid \text{i} \mid \text{li} \mid \text{a} \mid \text{zi} \mid \text{lu} \mid \text{bu} \mid \text{ku} \mid \epsilon$

$ns \rightarrow \text{angi} \mid \text{awu} \mid \text{aka} \mid \text{ali} \mid \text{asi} \mid \text{ayi} \mid \text{alu} \mid \text{abu} \mid \text{aku} \mid \text{ani} \mid$

$\text{aba} \mid \text{awa} \mid \text{azi} \mid \epsilon$

2. List of mod:

$m \rightarrow \text{a} \mid \text{e} \mid \text{ka} \mid \text{ma} \mid \text{nga} \mid \epsilon$

3. List of tense (nothing for the simple present tense):

$t \rightarrow \epsilon$

4. List of aspect (additional rules omitted in this first iteration):

$asp \rightarrow \text{sa} \mid \text{se} \mid \text{be} \mid \text{ile} \mid \epsilon$

5. List of object concords:

$o \rightarrow \text{ngi} \mid \text{si} \mid \text{ku} \mid \text{ni} \mid \text{m} \mid \text{ba} \mid \text{wu} \mid \text{yi} \mid \text{li} \mid \text{wa} \mid \text{zi} \mid \text{lu} \mid \text{bu} \mid \epsilon$

6. Causative:

- $c \rightarrow \text{is}$
7. Applicative:
 $a \rightarrow \text{el}$
8. Reciprocativative:
 $r \rightarrow \text{an}$
9. Passive (with phonological conditioning options):
 $p \rightarrow \text{iw} \mid \text{w}$
10. Lexicon of verb root:
 $vr \rightarrow \text{ab} \mid \dots \mid \text{zwib}$

This completes the first iteration: the core possibilities for present tense are completed with respect to R0 mentioned at the start of the section⁶. It can be optimised, but this is left for the implementation; here, we aimed to be as explicit as feasible.

4.2 Subsequent iterations

The outcome of the first iteration does not fully cover all verb options. Further extensions and refinements can be made, which are introduced now in their final version, being politeness, stative verbs, wh-questions, and aspect doubling.

Politeness The please and polite permissive questions have their own prefix system and a FV=e. This amounts to adding a new rule

$$ppre \rightarrow pl\ s$$

with the following terminals:

11. Please prefix, permissive prefix (none), and polite proposal doing something together, indicated with *pl*:

$$pl \rightarrow \text{aw} \mid \text{awu} \mid \text{mawu} \mid \varepsilon \mid \text{ma}$$

and extending the grammar rule for *v* with the extra option:

$$v \rightarrow pre\ vr\ post\ a \mid npre\ vr\ post\ i \mid ppre\ vr\ e$$

Stative verbs The stative refers to the state of being of something; e.g. *vula* ('open') with its stative variant *vuleka* ('be opened'), and *mbula* ('reveal') results in *mbuleka* ('be revealed'). This insertion of the *-ek-* between the VR and the FV is also referred to as the neuter extension. As it is conceptually different from the extension (i.e., CARP), we create a separate *st* and update the rule for *v* with it:

$$v \rightarrow pre\ vr\ post\ a \mid npre\ vr\ post\ i \mid ppre\ vr\ e \mid vr\ st\ a$$

with the following single terminal:

12. Stative verb, indicated with *st*:

$$st \rightarrow \text{ek}$$

Because there is only one terminal for *st*, the “*vr st a*”-part of *v* can also be written as “*vr eka*”.

⁶ Except that it does not take into account the swapping with OC and SC in case of P

Wh-questions The optional wh-questions fall in the post-final slot (see Table 1) and are added at the end of the verb, being *-ni* ‘what’/‘who’/‘why’/‘how’, *-nini* ‘when’, and *-phi* ‘where’. We create a separate *wh* variable for them and update the rule for *v* with it:

$$v \rightarrow pre\ vr\ post\ a\ wh \mid npre\ vr\ post\ i\ wh \mid ppre\ vr\ e \mid vr\ st\ a$$

with the following terminals for the new variable:

13. Wh-questions, indicated with *wh*:

$$wh \rightarrow ni \mid nini \mid phi \mid \varepsilon$$

Aspect doubling What is normally referred to as aspect doubling is a construction of aspect with continuous tense, i.e., the ‘second aspect’ is not an aspect in the strict sense of the meaning of aspect. Decomposed, we have EXCL-SC-CONT-(OC-)VR-(post)-a, where the exclusive can only be *se-* and continuous tense only *-ya-*. Because it is a regular exception, we add another ‘or’ to *v* rather than complicate *pre*:

$$v \rightarrow pre\ vr\ post\ a\ wh \mid npre\ vr\ post\ i\ wh \mid ppre\ vr\ e \mid vr\ st\ a \mid \\ excl\ s\ cont\ o\ vr\ post\ a$$

with the following terminals for the new variables:

14. ‘Double aspect’, indicated with *excl* for exclusive (with $excl \subset asp$)

$$excl \rightarrow se$$

15. With $cont \subset t$ and *cont* for continuous tense:

$$cont \rightarrow ya$$

16. The previous extension implies that *t* (item 3, above) also has to be updated:

$$t \rightarrow ya \mid \varepsilon$$

Finally, there is only one terminal for each, so the “*excl s cont o vr post a*”-part of *v* can also be written as “*se s ya o vr post a*”.

4.3 Other rules

While the CFG may seem alike a relatively free combination of anything, there are several constraints that are not covered by these grammar rules, as they would obfuscate the general patterns, not all of them are linguistically accounted for, and they are easier to implement as separate rules. Notably, there is an interaction between the two sides of the VR, which is ruled by the semantics of the CARP extension. For instance, for a construction to be causative and applicative, there have to be at least two things involved. The first participant is already catered for with the SC, the second is catered for with the OC. Typically, the causative and applicative will have an OC but the Reciprocal, Passive and the Stative would not. Further, for the passive, the object moves to the subject position, and so also with SC and OC.

The following set of rules (in pseudocode-style notation) is a first attempt at specifying them, and more will be added in due course:

- a) only **if** $p \in post$, **then: if** pre **then** $s \rightarrow \varepsilon$, **else** $ns \rightarrow \varepsilon$
- b) **if** $c \in post$, **then** $s, o \in pre$ or $npre$
- c) **if** $a \in post$, **then** $s, o \in pre$ or $npre$
- d) **if** $p \in post$, **then** $o \in (pre$ or $npre)$ and $s \notin (pre$ or $npre)$

- e) **if** $vr \in \text{Intransitive}$, **then** $r \notin \text{post}$ and $o \notin (\text{pre or } \text{npre})$
- f) **if** $vr \in \text{Monosyllabic}$, **then** $\text{post} \rightarrow \varepsilon$

The second set of rules have to do with phonological and morphological conditioning, such as:

- g) **if** $s==u$ **then** $pl=aw$, **else** $pl=aw$ or $mawu$ or ma

which we consider orthogonal to coverage of the different elements of the verb, and is therefore left for further work.

4.4 Extensions and other considerations

While the ‘other rules’ indicate intricate interactions between the various elements of a verb that might be addressed either with extra-CFG rules or a blow-up of CFG rules (by splitting the current ones in various ways) once fully known, one of a different kind is the treatment of the elements themselves. For instance, TAM can be at the start of the verb and at the end, but when at the start or end, only a *subset* of TAM is permissible, i.e., $FV \subset ? \subset TAM$, where that subset “?” exists, but is not well-documented yet as to why, what, and how.

The formal approach taken in the previous section lends itself well to a rigorous assessment of measurable distance or difference with verbs in other Bantu languages, as well as bootstrapping a CFG for some of the closely related but even lesser-resourced languages, such as Ndebele and isiXhosa. That said, we are well aware it will not be the same. Take, for instance, the Chishona—a neighbouring language—example in V5.

(V5) *mukomana a-ri-ku-gur-ir-a-zve* *chisikana*
 1.boy 1.SC-T-M-break-APPL-FV-too 7.girl
 ‘The boy is breaking (something) for the girl too’

The order of the extensions and clitics in the example above is worth noting. The clitic *-zve* comes after the FV *-a*. While this phenomenon is not found in isiZulu verb complex, it shows that there are unique features of the Bantu verb that further complicate the grammar, which will need to be accounted for.

Finally, incorporation of phonological and morphological conditioning, while being an orthogonal aspect to the structure of the components of the verb and order thereof, may, for practical reasons, have an effect on the rules itself. For instance, possibly splitting the terminals of the *vr* into one set for vowel-commencing roots and one for consonant-commencing roots, and then for ease of processing, some of the terminals of the *pre* and *npre* could be split into two as well.

5 Evaluation of the grammar

We first illustrate manually the functioning of the CFG with three use cases, and subsequently test it systematically with a computational version of it.

5.1 Use cases

Three examples are selected that also give a hint toward the CFG’s usability for a range of applications: generation of a word from the grammar (useful for machine translation and controlled natural languages), the checking of a correct word whether it is in the language of the grammar (spellchecking), and one misspelled word that gets rejected (correcting).

Let us step through the grammar in the least amount of steps (least amount of components) to ‘generate’ a word in the language, where each numbered subscript of the arrow is added for explanatory purpose afterward: $v \Rightarrow_1 pre\ vr\ post\ a\ wh \Rightarrow_2 s\ vr\ post\ a\ wh \Rightarrow_3 ngi\ vr\ post\ a\ wh \Rightarrow_4 ngi\ vel\ post\ a\ wh \Rightarrow_5 ngi\ vel\ a\ wh \Rightarrow_6 ngi\ vel\ a$; 1) substitute v for the first option; 2) substitute pre for the first option (s); 3) substitute s with the first terminal (**ngi**); 4) take one of the vrs (**vel**); 5) process $post$, choosing empty (ε); 5) process wh , choosing empty (ε). Thus, the word generated is: **ngivela** ‘I come from’.

Stepping through the grammar, using more slots at the end, we can check that **niboniselana** is in the language: $v \Rightarrow pre\ vr\ post\ a\ wh \Rightarrow ni\ vr\ post\ a\ wh \Rightarrow ni\ bon\ post\ a\ wh \Rightarrow ni\ bon\ c\ a\ r\ a\ wh \Rightarrow ni\ bon\ is\ a\ r\ a\ wh \Rightarrow ni\ bon\ is\ el\ r\ a\ wh \Rightarrow ni\ bon\ is\ el\ an\ a\ wh \Rightarrow ni\ bon\ is\ el\ an\ a$.

The grammar thus also can be used to recognising misspelled words. For instance, a user types ***usafundapi**, then it rejects at the **-pi** end: $v \Rightarrow pre\ vr\ post\ a \Rightarrow s\ asp\ vr\ post\ a\ wh \Rightarrow u\ asp\ vr\ post\ a\ wh \Rightarrow u\ sa\ vr\ post\ a\ wh \Rightarrow u\ sa\ fund\ post\ a\ wh \Rightarrow u\ sa\ fund\ a\ wh \Rightarrow \times$. The trace/tree can not be completed because **pi** \notin **wh** (**phi** and **ni** are), thus ***usafundapi** is misspelled with respect to the grammar rules as introduced in the previous sections. Proposing a correction can be done by suggesting to complete **usafunda-** with any of the **wh** terminals, or, when using the minimum edit distance as an extra service in the spellchecker, it would suggest **usafundaphi** ‘where are you still studying?’ and **usafundani** ‘what/why are you still studying?’ as the two options to choose from to correct the misspelled word.

5.2 Computational evaluation of the grammar

There are many tools that are candidates to implement the grammar to the point of testing whether the rules are the right ones; that is, the scope is *validation* (‘are we building the right grammar?’) and *verification* (‘are we building the grammar right?’), not end-user tool building.

Implementation considerations Most computational linguistics papers for Bantu languages use one of the tools for building a morphological analyser. Xfst and lexc has been used to encode a subset of the rules for verbs in isiZulu, Setswana, and Ekegusii [22,21,8], whereas Fsm2 could not be found online anymore. However, they are problematic theoretically. Xfst, and similar tools such as SFST⁷ and OpenFST⁸, are transducers, and therewith limited to regular gram-

⁷ <http://www.cis.uni-muenchen.de/~schmid/tools/SFST/>

⁸ <http://www.openfst.org/twiki/bin/view/FST/WebHome>

mars (The surface syntax gives the impression of accepting a CFG, but that is syntactic sugar and is transformed behind-the-scenes into a (very) large FSA). While most of the rules in the previous section look indeed regular, for being in a fixed order at least, when $p \in post$, then the $o \in pre$ takes the position of the s . This already indicates that the grammar for the verb on its own is beyond a regular grammar, hence, beyond a FSM, so a transducer is insufficiently expressive. This is unsurprising, as natural languages tend all to be context-free [25]. Another option is to take a programmatic approach. Python programming language is popular, used by [27], and the NLTK [2] has a CFG grammar module. However, the latter requires the word already to be segmented, but this is precisely what needs to happen automatically, and building a regular expression grammar faces the same issue as mentioned above. Spiegler et al’s DCG for the Ukwabelana tagging is in Prolog, but at this validation and verification stage a full-fledged tool is not needed. Therefore, we used the JFlap tool⁹, which can check string membership and generate words in the language.

Testing in JFlap Transferring the written grammar into JFlap (v8 beta) ironed out two glitches in variable abbreviations (corrected version is included in the previous section), and some of the variable names are different, because the tool allows only single-character variable names. The JFlap file, conversion annotations, and the screenshots of the outputs are available online at <http://www.meteck.org/files/geni/>.

We selected a set of verbs that covers the principal permutations of the rules, and some that ought to be rejected, as indicated in Table 4. The strings in the first set were all accepted; a screenshot of the derivation table of **niboniselana** is shown in Fig. 3 and the screenshots for the others are in the online material. Thus, the CFG recognises what it should recognise, and thus indicates correctness of the grammar specified. Of the terms one would have liked to have it rejected, only **ngiveli** was (incorrectly) accepted, which is due to the ε ’s that are in the grammar due to the absence of the extra rules in the JFlap CFG (recall Section 4.3), so $npre$ is decomposed as $ns\ o$, with $ns \Rightarrow \varepsilon$ and $o \Rightarrow \mathbf{ngi}$. Thus, the strings that the grammar accepts/generates are more words than that are in the isiZulu language. This can be seen also with the ‘generate strings’ feature in JFlap. Setting the number of strings to a subset, 100, to check, showed that this is due to not only not including the additional constraints but also not catering yet for phonological conditioning; e.g., the aforementioned ***ngiveli** and ***aabaphi**, where **a**=SC nc:6, **a** = Mood, **ba** = VR ‘distribute’, and **phi** = wh ‘where’, but it requires an consonant between the two a’s. Addressing this issue is orthogonal to the grammar, and therefore left for future work.

Further, we checked our grammar against the Zulu finite state morphology demo of the Academy of African Languages and Science¹⁰. As one may expect, it accepts **niboniselana**, but it also accepts ***nibonelisana**, which has the wrong

⁹ <http://www.cs.duke.edu/csed/jflap/>

¹⁰ <http://gama.unisa.ac.za/demo/demo/zulmorph>; tested with the version online d.d. 17-12-2015.

Table 4. Strings selected for testing; A/R: accepted/rejectedd.

String	Reason	A/R	Correct
ngivela	simple present tense	A	+
angiveli	simple negation	A	+
angivela	<i>pre</i> with <i>s</i> and <i>o</i>	A	+
asingabaveli	testing <i>npre</i>	A	+
niboniselana	testing <i>post</i>	A	+
vuleka	stative	A	+
usafundaphi	wh-extension	A	+
sengiyafunda	aspect doubling	A	+
awusidle	politeness	A	+
*ngiveli	mixing positive with negative FV	A	–
*usafundapi	typo; wrong wh-extension	R	+
*nibonelisana	wrong CARP order	R	+
*sangiyafunda	wrong aspect in aspect doubling	R	+
*kabevela	wrong order in <i>pre</i> , no <i>sc</i>	R	+

CARP order, and accepts *kabevela (MOD-ASP-VR-FV), which has MOD and ASP in the wrong order (and lacks SC/OC) and is therefore rejected by our CFG. That is, that FSM does not handle any order of components of the verb.

While using a CFG computationally is an error-proof method for ‘finding’ a derivation, the CFG up to the wh-extension used up 773240 nodes in the brute force parser to accept **niboniselana**, due to exploring all potential paths to completion. This was with five verb roots for testing the grammar; adding another five verb roots generated 1086109 nodes in order to accept **niboniselana**. With the aspect doubling extension, this increased further to 1168099 nodes; see Figure 4. Computationally, this is clearly not sustainable with a brute force parser, and a practical implementation that uses the CYK algorithm instead will be needed. Nevertheless, the brute-force parser is useful for evaluating the grammar.

6 Conclusions

We presented the precise specification of the isiZulu verb present tense as a Context-Free Grammar. It covers not only the usual subject and object concords, but also negation, present tense, aspect, mood, and the verbal extensions such as the causative, applicative, stative and the reciprocal, politeness, the wh-questions modifiers, and aspect doubling, all in their correct order as they appear in verbs. In addition to a paper-based specification, it was represented computationally as a CFG in the JFlap tool and tested on correctness of specification, using a set of words and generating their derivations in the JFlap tool. The grammar conforms to specification, though still accepts more strings than those that are in the isiZulu language. This is due to the absence of additional rules in the

Derivation Tree Derivation Table	
Production	Derivation
	V
V->A R B a N	A R B a N
A->S	S R B a N
S->n i	n i R B a N
R->b o n	n i b o n B a N
B->F G H	n i b o n F G H a N
F->i s	n i b o n i s G H a N
G->e l	n i b o n i s e l H a N
H->a n	n i b o n i s e l a n a N
N-> λ	n i b o n i s e l a n a

Fig. 3. Screenshot of the derivation table as computed by JFlap (brute force parser), on **niboniselana**.

implementation and the orthogonal issue of phonological conditioning, which are aspects of future work.

Acknowledgements This work is based on the research supported in part by the National Research Foundation of South Africa (CMK: Grant Number 93397).

References

1. Beuchat, P.D.: The verb in Zulu. Johannesburg: Witwatersrand University Press (1966), reprinted from African Studies 22: 137-169, 1963; 23: 35-49, 67-87, 1964; 25: 61-71, 1966
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
3. Bosch, S.E., Eiselen, R.: The effectiveness of morphological rules for an isiZulu spelling checker. South African Journal of African Languages 25(1), 25–36 (2005)
4. Buell, L.C.: Issues in Zulu Verbal Morphosyntax. Ph.d. thesis, University of California, Los Angeles (2005)
5. Doke, C.: Text Book of Zulu Grammar. Witwatersrand University Press (1927)
6. Doke, C.: Bantu Linguistic Terminology. London: Longman, Green and Co (1935)
7. Durrant, P.: Formulaicity in an agglutinating language: the case of Turkish. Corpus Linguistics and Linguistic Theory 9(1), 1–38 (2013)
8. Elwell, R.: Finite-state Methods for Bantu Verb Morphology. In: Gaylord, N., Hilderbrand, S., Lyu, H., Palmer, A., Ponvert, E. (eds.) Texas Linguistics Society 10: Computational Linguistics for Less-Studied Languages. CSLI Publications (2005)
9. Getao, K., Miriti, E.: Special Topics in Computing and ICT Research: Computational modelling in Bantu language. Advances in Systems Modelling and ICT Applications pp. 128–138 (2000)
10. Guthrie, M.: Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages. No. v. 1-2, Gregg (1971)

Level	Total Nodes	Current Derivations
1	5	[A R B a N, C R B i N, E S Q O R B a, J R e, R L a]
2	131	[R B a N, S R B a N, S M R B a N, S M O R B a N, S ...]
3	1721	[L a B a N, R a N, R F G a N, R F G H a N, R F G I a ...]
4	13321	[L a F G a N, L a F G H a N, L a F G I a N, L a F H a ...]
5	64421	[L a F e l a N, L a F G a, L a F e l H a N, L a F G a n ...]
6	221137	[L a F e l a n a N, L a F e l H a, L a F G a n a, L a F ...]
7	586940	[L a F e l a n a, R i s e l a n a, n i L a i s e l a n, n i ...]
8	1168074	[n i L a i s e l a n a N, n i L a i s e l H a, n i L a i s ...]
9	1168099	[niboniselana]

Fig. 4. Screenshot of JFlap’s output with the brute force parser for the final CFG, on niboniselana.

11. Hyman, L.: Conceptual issues in the comparative study of the bantu verb stem. In: Mufwene, S.S., Moshi, L. (eds.) *Topics in African Linguistics*, pp. 3–34. Amsterdam/Philadelphia, John Benjamins Publishing Co. (1991)
12. Katushemerewe, F., Hanneforth, T.: Finite state methods in morphological analysis of Runyakitara verbs. *Nordic Journal of African Studies* 19(1), 1–22 (2010)
13. Keet, C.M., Khumalo, L.: Basics for a grammar engine to verbalize logical theories in isiZulu. In: Bikakis, A., et al. (eds.) *Proceedings of the 8th International Web Rule Symposium (RuleML’14)*. LNCS, vol. 8620, pp. 216–225. Springer (2014), august 18–20, 2014, Prague, Czech Republic
14. Keet, C.M., Khumalo, L.: Toward verbalizing logical theories in isiZulu. In: Davis, B., Kuhn, T., Kaljurand, K. (eds.) *Proceedings of the 4th Workshop on Controlled Natural Language (CNL’14)*. LNAI, vol. 8625, pp. 78–89. Springer (2014), 20–22 August 2014, Galway, Ireland
15. Khumalo, L.: An analysis of the Ndebele Passive Construction. Ph.D. thesis, University of Oslo, Norway (2007)
16. Lindfors, A.L.: Tense and aspect in swahili. Technical report, Institutionen for Lingvistik, Uppsalla Universitet (2003), http://www2.lingfil.uu.se/ling/semfiler/Swa_TAM.pdf
17. Mabugu, P.: Polysemy and the Applicative Verb Construction in Chishona. Doctoral dissertation, University of Edinburgh (2001)
18. Mchombo, S.: The syntax of Chichewa. Cambridge, UK: Cambridge University Press (2004)
19. Mchombo, S.: Argument binding and morphology in chichewa. In: Hoyt, F., Seifert, N., Teoderescu, A., White, J. (eds.) *Texas Linguistics Society 9: Morphosyntax of Underrepresented Languages*, pp. 203–221. Texas Linguistics Society (2007)
20. Meeussen, A.E.: Bantu grammatical reconstructions. *Africana Linguistica* 3, 79–121 (1967)
21. Pretorius, R., Berg, A., Pretorius, L., Viljoen, B.: Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography.

- In: Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages (AfLaT'09). pp. 66–73 (2009), athens, Greece, 31 March 2009
22. Pretorius, L., Bosch, E.S.: Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation* 18(3), 195–216 (2003)
 23. Pretorius, L., Bosch, S.E.: Finite state morphology of the Nguni language cluster: modelling and implementation issues. In: Yli-Jyrä, A., Kornai, A., Sakarovitch, J., Watson, B. (eds.) *Finite-State Methods and Natural Language Processing 8th International Workshop (FSMNLP'09)*. LNCS, vol. 6062, pp. 123–130. Springer (2009)
 24. Pretorius, L., Bosch, S.: Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele. In: *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*. pp. 73–78 (2012)
 25. Pullum, G.K., Gazdar, G.: Natural languages and context-free languages. *Linguistics and Philosophy* 4(4), 471–504 (1982)
 26. Schadeberg, T.C.: Derivation. In: Nurse, D., Philippson, G. (eds.) *The Bantu Languages*. Routledge (2003)
 27. Spiegler, S., van der Spuy, A., Flach, P.A.: Ukwabelana – an open-source morphological Zulu corpus. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*. pp. 1020–1028. Association for Computational Linguistics (2010), beijing
 28. Wald, B.: Swahili and the Bantu languages. In: Comrie, B. (ed.) *The World's Major Languages*, pp. 991–1014. Oxford: Oxford University Press (1987)
 29. Zawawi, S.M.: *Loan words and their effect on the classification of Swahili nominals*. Leiden: E. J. Brill (1979)